# Structured Errors in Optical Gigabit Ethernet Packets

Laura James, Andrew Moore, Madeleine Glick

IRC –TR-04-021

April 2004

*PAM*

DISCLAIMER: THIS DOCUMENT IS PROVIDED TO YOU "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY NON-INFRINGEMENT, OR FITNESS FOR ANY PARTICULAR PURPOSE. INTEL AND THE AUTHORS OF THIS DOCUMENT DISCLAIM ALL LIABILITY, INCLUDING LIABILITY FOR INFRINGEMENT OF ANY PROPRIETARY RIGHTS, RELATING TO USE OR IMPLEMENTATION OF INFORMATION IN THIS DOCUMENT. THE PROVISION OF THIS DOCUMENT TO YOU DOES NOT PROVIDE YOU WITH ANY LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS

Copyright 2004, Intel Corporation, All rights reserved.

**Intel Research Cambridge**

# Structured Errors in Optical Gigabit Ethernet Packets

Laura James[1], Andrew Moore[2], and
Madeleine Glick[3]

[1] Centre for Photonic Systems, University of Cambridge
lbj20@eng.cam.ac.uk
[2] Computer Laboratory, University of Cambridge
andrew.moore@cl.cam.ac.uk
[3] Intel Research Cambridge
madeleine.glick@intel.com

**Abstract.** This paper presents a study of the errors observed when an optical Gigabit Ethernet link is subject to attenuation. We use a set of purpose-built tools which allows us to examine the errors observed on a per-octet basis. We find that some octets suffer from far higher probability of error than others, and that the distribution of errors varies depending on the type of packet transmitted.

## 1 Introduction

Optical communications assessment is conventionally based upon Bit Error Rate (BER), a purely statistical measurement. Higher level network design decisions are often based on an assumption of uniform error at the physical layer, with errors occurring independently and with equal probability regardless of data value or position.

We examine the assumption of uniform behaviour in Gigabit Ethernet at low optical power, and find that the failure mode observed is non-uniform, caused by interactions between the physical layer, the coding system and the data being carried.

### 1.1 Motivation

The assumed operating environment of the underlying coding scheme must be re-examined as new more complex optical systems are developed. In these systems containing longer runs of fibre, splitters, and active optical devices, the overall system loss will be greater than in today's point-to-point links, and the receivers may have to cope with much lower optical powers. Increased fibre lengths used to deliver Ethernet services, e.g. the push for Ethernet in the first mile [1], and switched optical networks [2] are examples of this trend.

The design of the physical and data link layers affects how resilient a network is to errors in the communications medium. Understanding the behaviour

of the medium and probable error types and distributions, is necessary to design a network that will avoid error causes, and compensate for others. Coding schemes such as 8B/10B have many desirable properties, and therefore a thorough understanding of the effects of error is important for future optical network design.

### 1.2 Gigabit Ethernet Physical Coding Sublayer

Using a transmission code improves the resilience of a communications link, by ensuring the data stream has known characteristics that are well matched to the physical behaviour of the link. A coding scheme must ensure the recovery of transmitted bits; often this requires a minimum number of bit transitions to occur for successful clock and data recovery. In most systems, transceivers are AC coupled, which can lead to distorted pulses and baseline wander (as the DC component of the signal builds up); these can be reduced by the use of a block code which is symmetrical around the zero line.

The 8B/10B codec, originally described by Widmer & Franaszek [3] sees use in 1000BASE-X, optical Ethernet. This scheme converts 8 bits of data for transmission (ideal for any octet-orientated system) into a 10 bit line code. Although this adds a 25% overhead, 8B/10B has many valuable properties; a transition density of at least 3 per 10 bit code group and a maximum run length of 5 bits for clock recovery, along with virtually no DC spectral component. These characteristics also reduce the possible signal damage due to jitter, which is particularly critical in optical systems, and can also reduce multimodal noise in multimode fibre connections.

This coding scheme is royalty-free and well understood, and is currently used in a wide range of applications; in addition to being the standard for optical gigabit Ethernet, it is used in the Fibre Channel system [4], and 8B/10B coding will be the basis of coding for the electrical signals of the upcoming PCI Express standard [5].

**8B/10B Coding** The 8B/10B codec defines encodings for data octets, and control codes which are used to delimit the data sections and maintain the link. Individual codes or combinations of codes are defined for Start of Packet, End of Packet, line Configuration, and so on. Also, Idle codes are transmitted when there is no data to be sent, and these keep the transceiver optics and electronics active. The Physical Coding Sublayer (PCS) of the Gigabit Ethernet specification [6] defines how these various codes are used.

Individual ten bit code-groups are constructed from the groups generated by 5B/6B and 3B/4B coding on the first five and last three bits of a data octet respectively. Some examples are given in Table 1; the running disparity is the sign of the running sum of the code bits, where a one is counted as 1 and a zero as -1. During an Idle sequence between packet transmissions, the running disparity is changed (if necessary) to -1, and then maintained at that value. Both control and data codes may change the running disparity, or may preserve its existing

value; examples of both types are shown in Table 1. The code-group used for the transmission of an octet depends upon the existing running disparity – hence the two alternative codes given in the table. A received code-group is compared against the set of valid code-groups for the current receiver running disparity, and decoded to the corresponding octet if it is found. If the received code is not found in that set, the specification states that the group is deemed invalid. In either case, the received code-group is used to calculate a new value for the running disparity. A code-group received containing errors may thus be decoded and considered valid. It is also possible for an earlier error to throw off the running disparity calculation, such that a later code-group may be deemed invalid, as the running disparity at the receiver is no longer correct. This can amplify the effect of a single bit error at the physical layer. Line coding schemes, although

**Table 1.** Examples of 8B/10B control and data codes

| Type | Octet | Octet bits | Current RD - | Current RD + | Note |
|---|---|---|---|---|---|
| data | 0x00 | 000 00000 | 100111 0100 | 011000 1011 | preserves RD value |
| data | 0xf2 | 111 10010 | 010011 0111 | 010011 0001 | swaps RD value |
| control | K27.7 | 111 11011 | 110110 1000 | 001001 0111 | preserves RD value |
| control | K28.5 | 101 11100 | 001111 1010 | 110000 0101 | swaps RD value |

they handle many of the physical layer constraints, can introduce problems. In the case of 8B/10B coding, a single bit error on the line can lead to multiple bit errors in the received data byte. For example, with one bit error the code-group D0.1 (current running disparity negative) becomes the code-group D9.1 (also negative disparity); these decode to give bytes with 4 bits of difference. In addition, the running disparity after the code-group is miscalculated, potentially leading to future errors. There are other similar examples [6].

## 2 Method

In this paper we investigate Gigabit Ethernet on optical fibre, (1000BASE-X [6]) when the receiver power is sufficiently low as to induce errors in the Ethernet frames. We assume that while the CRC mechanism within Ethernet is sufficiently strong as to catch the errors, the dropped frame and resulting packet loss will result in hosts, applications and perhaps users whose packets will be in error with a significantly higher probability than the norm.

In our main test environment an optical attenuator is placed in one direction of a Gigabit Ethernet link. A traffic generator feeds a Fast Ethernet link to an Ethernet switch, and a Gigabit Ethernet link is connected between this switch and a traffic sink and tester. The variable optical attenuator is placed in the fibre in the direction from the switch to the sink.

A packet capture and measurement system is implemented within the traffic sink using an enhanced driver for the SysKonnect SK-9844 network interface card (NIC). Among a number of additional features, the modified driver allows application processes to receive frames that contain errors that would normally be discarded. Alongside purpose-built code for the receiving system we use a special-purpose traffic generator. Pre-constructed test data in tcpdump-format file is transmitted from one or more traffic generators using *tcpfire* [7]. By transmitting a pre-determined traffic stream we can identify specific errored octets within the received data stream.

### 2.1 Octet Analysis

Each octet for transmission is coded using 8B/10B into a 10 bit code group or *symbol*, and we analyze these for frames which are received in error at the octet level. By comparing the two possible transmitted symbols for each octet in the original frame, to the two possible symbols corresponding to the received octet, we can deduce the bit errors which occurred in the symbol at the physical layer. In order to infer which symbol was sent and which received, we assume that the combination giving the minimum number of bit errors on the line is most likely to have occurred. This allows us to determine the line errors which most probably occurred.

Various types of symbol damage may be observed. One of these is the single bit error caused by the low signal to noise ratio at the receiver. A second form of error results from a loss of bit clock causing *smeared* bits: where a subsequent bit is read as having the value of the previous bit. A final example results from the loss of symbol clock synchronization. This can result in the symbol boundaries being misplaced, so that a sequence of several symbols, and thus several octets, will be incorrectly recovered.

### 2.2 Real Traffic

Some experiments are conducted with real network traffic referred to as the *day-trace*. This network traffic was captured from the interconnect between a large University department and that universities' principle data-backbone over the course of two working days. We consider it to contain a representative sample of network traffic for an academic/research department network with approximately 150 users.

The probability of occurrence of each octet value within the day-trace is given in Figure 1. The illustrated probabilities allow specific insight into the effect of symbol-errors within a realistic traffic workload.

Other traffic tested included *pseudo-random data*, consisting of a series of 1500 octet frames each filled with octets whose values were each drawn from a pseudo-random number generator. *Structured test data* consists of a single frame containing repeated octets: 0x00–0xff, to make a frame 1500 octets long. The *low error testframe* consists of 1500 octets of 0xCC data (selected for a low symbol

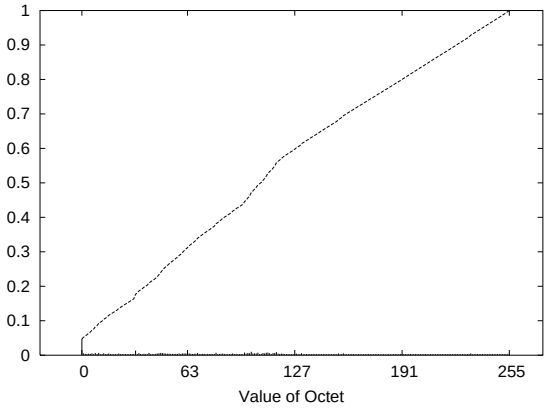


**Fig. 1.** Probability of occurrence for a particular octet within *daytrace*

error rate); the *high error testframe* is 1500 octets of 0x34 data (which displays a high symbol error rate).

### 2.3 Bit Error Rate Measurements

In photonics laboratories, a Bit Error Rate Test kit (BERT) is commonly used to assess the quality of an optical system [8]. This comprises both a data source and a receiving unit, which compares the incoming signal with the known transmitted one. Individual bit errors are counted both during short sampling periods and over a fixed period (defined in time, bits, or errors). The output data can be used to modulate a laser, and a photodiode can be placed before the BERT receiver to set up an optical link; optical system components can then be placed between the two parts of the BERT for testing. Usually, a pseudo-random bit sequence is used; but any defined bit sequence can transmitted repeatedly and the error count examined.

For the bit error rate measurements presented here, a directly modulated 1548nm laser was used (the wavelength of 1000BASE-ZX). The optical signal was then subjected to variable attenuation, before returning via an Agilent Lightwave (11982A) receiver unit into the BERT (Agilent parts 70841B and 70842B). The BERT was programmed with a series of bit sequences, each corresponding to a frame of Gigabit Ethernet data, encoded as it would be for the line in 8B/10B. Purpose-built code is able to construct the bit-sequence, suitable for the BERT, from a frame of known data. The bit error rates for these various packet bit sequences were measured at a range of attenuation values, using identical BERT settings for all frames (eg. 0/1 thresholding value).

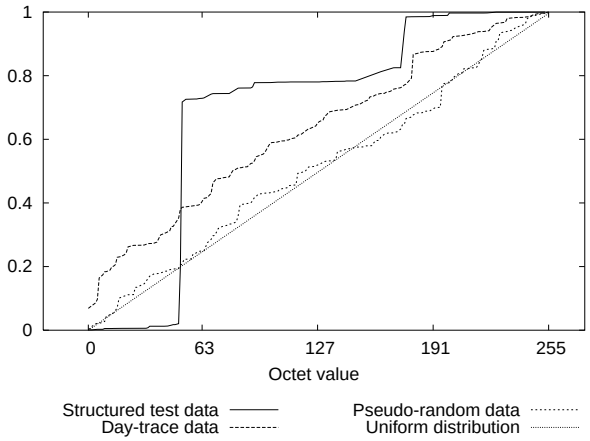


**Fig. 2.** Cumulative distribution of errors versus octet value

## 3 Results

A plot of the cumulative distribution of errors for three traffic types is shown in Figure 2. Note that while the random data frames closely follow the expected uniform error distribution, the *day-trace* frames suffer from higher error rates in the low value octets, especially the value 0x00. The structured test data shows an even more significant error rate focused upon only a small number of octets (e.g., 0x34). The test attenuation used here corresponds to a frame loss of 892 in $10^6$ pkts for structured test data, to 233 in $10^6$ pkts for pseudo-random data, and to 98 in $10^6$ pkts for the *day-trace* data.

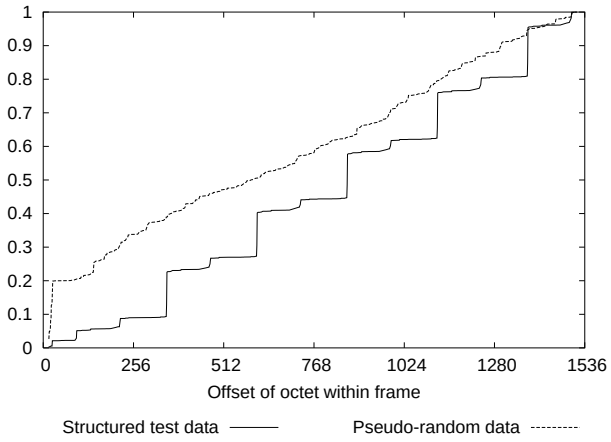


**Fig. 3.** Cumulative distribution of errors versus position in frame

Figure 3 contrasts the position of errors in the pseudo-random and structured frames. The *day-trace* results are not shown as this data, consisting of packets of varying length, is not directly comparable. The positions of the symbols most subject to error can be clearly observed. In contrast with the structured data, the random data shows a much increased error rate at the beginning of the frame.

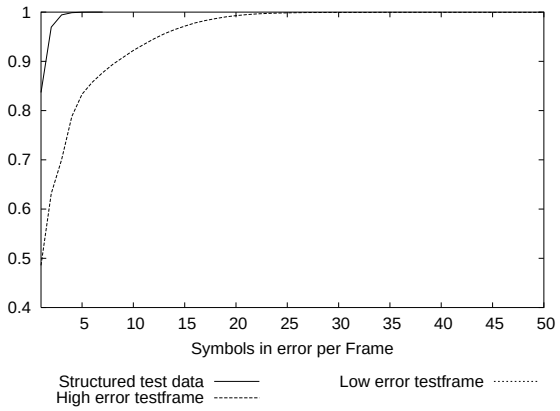


**Fig. 4.** Cumulative distribution of symbol errors per frame

Figure 4 indicates the number of incorrectly received symbols per frame. The *low error testframe* generated no symbols in error, and thus no frames in error. The *high error testframe* results show an increase in errored symbols per frame. It is important to note that the errors occur uniformly across the whole packet, and there are no correlations evident in the position of errors within the frame.

Figure 5 shows the frequency of received packets in error, for a range of receiver optical power and for the three different testframes.

Figure 6 shows the bit error rate, as measured using the BERT, for a range of receiver optical power. These powers are different to those in Figure 5 due to the different experimental setup; packet error is measured using the SysKonnect 1000BASE-X NIC as a receiver, and bit error with an Agilent Lightwave receiver, which have different sensitivities.

Figures 5 and 6 show clearly different error rates for the different frames transmitted. A frame giving unusually high bit error rates at the physical layer does not necessarily lead to high rates of packet error.

## 4 Discussion

From Figure 2, it can be seen that 0x00 data suffers from greater error than the assumed uniform distribution would suggest when transmitted as part of a real traffic stream. Given the high proportion of 0x00 octets in genuine traffic (Figure 1), this effect is likely to be amplified. When the position of errors within

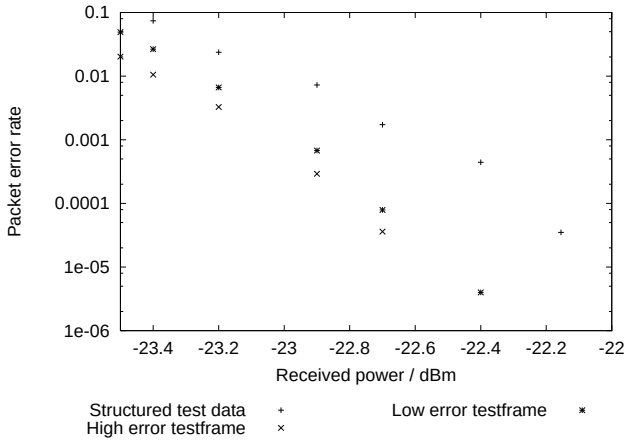


**Fig. 5.** Frequency of packet error versus received power

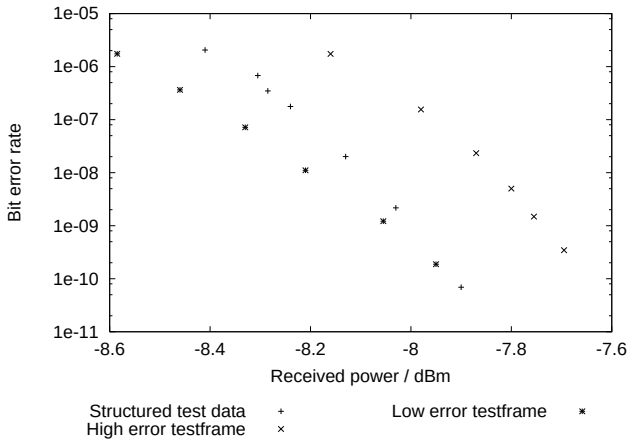


**Fig. 6.** Bit error rate versus received power

a frame is considered in Figure 3, we find that pseudo-random data is particularly subject to errors at the start of the frame (compared to the structured testframe, where errors occur at certain octets throughout the frame). If this behaviour also occurred in real frames, it may be the header fields which would be most subject to error. We consider example frames which consist of octets leading to high and low symbol error rates, and find that a higher symbol error probability leads to an increased number of errors per frame (Figure 4). Regardless of the pattern or data structure in which these octets are found in the frame, the individual likelihoods of their being received in error still have a noticeable effect.

When we compare the packet error rates obtained using our main test system with the bit error rates from the BERT measurements (Figures 5 and 6), we see that these different testframes lead to substantially different BER performance. Also, the BER is no indicator of the packet error probability.

We thus find that some octets suffer from a far higher probability of error than others, and that the distribution of errors varies depending on the type of packet transmitted. When the network carrying data is operating in a regime of limited power at the receiver, the initial implication of our work is that certain combinations of data will be subject to a significantly increased error rate. We conjecture that the PCS of Gigabit Ethernet can cause *hot-spots* to occur within the data symbols, hot-spots that contradict the underlying assumption of uniformity of error. The result is that, in our example, some Ethernet frames carried over 1000BASE-X upon the 8B/10B coding scheme will suffer a higher rate of error than might otherwise be assumed. More specifically, an Ethernet frame carrying certain octets, early in the payload, is up to 100 times more likely to be received with errors (and thus dropped), than if the payload does not contain such *hot-spot* octets. In addition, we observe that the structure of the packet can worsen this effect.

The error *hot-spots* we have observed may have implications for higher level network protocols. Frame-integrity checks, such as a cyclic redundancy check, assume that there will be a uniformity of errors within the frame, justifying detection of single-bit errors with a given precision. One example: Stone *et al.* [9] provides insight into such non-uniformity of data, discussing the impact this has for the checksum of TCP. These results may call into question our assumption that only increased packet-loss will be the result of the error *hot-spots*. Instead of just lost packets, Stone *et al.* noted certain "unlucky" data would rarely have errors detected. Another example of related work is Jain [10] which illustrates how errors will impact the PCS of FDDI and require specific error detection/recovery in the data-link layer.

In further related work, researchers have observed that up to 60% of faults in an ISP-grade network are due to optical-events [11]; while the majority of these will be catastrophic events, we would speculate that including the hot-spotting we observed, a higher still optically-induced fault rate will exist.

We highlight here an unexpected failure mode that occurs in the low-power regime, inducing at worst: errors, and at best: poor performance, that may focus upon specific networks, applications and users.

## 5 Conclusion

We have shown that the errors observed in Gigabit Ethernet in a low-power regime are not uniform as may be assumed, and that some packets will suffer greater loss rates than the norm. This content-specific effect will occur without a total failure of the network, and so must be given careful consideration. Even running the system above the power limit, with a BER of $10^{-12}$, a line error may occur as often as every 800 seconds. Our observation that real network traffic exhibited non-uniform symbol damage suggests that optical networks carrying real traffic will suffer hot-spots of higher frame loss rate than the physical layer conditions would indicate. Future work will clarify these effects.

### Acknowledgements

Many thanks to Derek McAuley, Richard Penty, Dick Plumb, Adrian P. Stephens, Ian White and Adrian Wonfor for their assistance, insight and feedback throughout the preparation of this paper. Laura James would like to thank Marconi Communications for their support of her research. Andrew Moore would like to thank Intel Corporation for their support of his research.

## References

1. IEEE: IEEE 802.3ah — Ethernet in the First Mile (2003) Task Force.
2. McAuley, D.: Optical Local Area Network. In Herbert, A., Spärck-Jones, K., eds.: Computer Systems: Theory, Technology and Applications. Springer-Verlag (2003)
3. Widmer, A.X., Franaszek, P.A.: A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code. IBM Journal of Research and Development **27** (1983) 440–451
4. The Fibre Channel Association: Fibre Channel Storage Area Networks. LLH Technology Publishing, Eagle Rock, VA (2001)
5. Solari, E., Congdon, B.: The Complete PCIExpress Reference. Intel Press, Hillsboro, OR (2003)
6. IEEE: IEEE 802.3z — Gigabit Ethernet (1998) Standard.
7. : tcpfire (2003) http://www.nprobe.org/tools/.
8. Ramaswami, R., Sivarajan, K.N. In: Optical Networks. Morgan Kaufmann (2002) 258–263
9. Stone, J., Greenwald, M., Partridge, C., Hughes, J.: Performance of Checksums and CRCs over Real Data. In: Proceedings of ACM SIGCOMM 2000, Stockholm, Sweden (2000)
10. Jain, R.: Error Characteristics of Fiber Distributed Data Interface (FDDI). IEEE Transactions on Communications **38** (1990) 1244–1252
11. Markopoulou, A., Iannaccone, G., Bhattacharyya, S., Chuah, C.N., Diot, C.: Characterization of failures in an IP backbone. In: To appear in IEEE Infocom, Hong Kong (2004)